基于深度强化学习的生鲜产品联合库存控制与动态定价研究 *

毕文杰，周玉冰

(中南大学 商学院，长沙 410083)

摘　要：针对由于生鲜产品的易逝性特征以及复杂多变的现实环境导致生鲜产品的最优订货和定价策略难以获得问题，提出了基于深度强化学习方法的生鲜产品联合库存控制与动态定价方法，结合生鲜产品特性对问题进行建模并定义为马尔可夫决策过程，然后基于深度强化学习设计了生鲜品联合库存控制和动态定价算法。实验结果表明，基于深度强化学习的联合库存控制和动态定价策略收益表现最佳，因此，基于深度强化学习的联合库存控制和动态定价研究能够提高企业收益，有效促进强化学习在收益管理领域的落地，具有实际应用价值。

关键词：深度强化学习；收益管理；生鲜产品；库存控制；动态定价

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2022.01.0056

Research on inventory control and dynamic pricing of fresh produce based on deep reinforcement learning

Bi Wenjie, Zhou Yubing

(*Business School*, *Central South University*, *Changsha* 410083, *China*)

Abstract: Due to the perishable characteristics of fresh produce and the complex and changing environment, it is difficult to obtain the optimal ordering and pricing strategy for fresh produce. To solve this problem, this paper proposed a deep reinforcement learning method for joint inventory control and dynamic pricing of fresh produce. The method combined the characteristics of fresh produce to model the problem and defined it as a Markov decision process. Then, this paper designed a joint inventory control and dynamic pricing algorithm for fresh produce based on deep reinforcement learning. The experimental results showed that the inventory control and dynamic pricing algorithm designed based on deep reinforcement learning had the best performance in terms of revenue. Therefore, the research on joint inventory control and dynamic pricing based on deep reinforcement learning methods can effectively improve enterprise revenue and promote the implementation of reinforcement learning in the field of revenue management, which has practical application value.

Key words: deep reinforcement learning; revenue management; fresh produce; inventory control; dynamic pricing

0　引言

近年来，人们物质生活水平显著提高，日益追求品质生活，消费者对生鲜产品的需求不断增长。而由于电商的迅猛发展以及物流系统的不断完善，消费者对产品品质要求也越发严格。根据国家统计局 2020 年鉴统计数据显示，生鲜产品在全国居民人均主要食品消费量中的占比超过一半且需求量呈现逐年递增的趋势。

由于生鲜产品具有广阔的市场前景，并且近年电子商务与冷链物流不断发展，“生鲜网购”模式兴起。企业在互联网上销售新鲜蔬果和生鲜肉类，发展出了“到店”、“到店+到家”、社区团购、“到柜”等多种多样的经营模式，生鲜产品极具市场潜力。但 2018 年全国生鲜电商只有 1%实现了盈利；2019 年更是众多中小企业纷纷关店关停；2020 年受疫情影响，生鲜电商迎来了行业的春天，但同时竞争加剧。因此，生鲜电商亟需寻求解决方案以突破困境。

同时，考虑到生鲜产品属于易逝品，表现出明显的易逝性：时令性、生命周期短、期末未售出产品的残值低以及需求随机性强等特征[1,2]。这些特性导致产品本身易损腐，销售风险增高。因此，企业需要考虑到产品的易逝性，制定合理的库存及定价策略，才能实现收益最大化。除此之外，产品的价格会影响需求，进而影响到最佳库存策略。这种价格和库存相互依存的关系意味着定价和库存控制应该同时进行决策，联合决策也能有效提升企业收益。Fang[3]发现，相比仅优化库存策略，联合定价和库存决策可以令高级时装企业的收益提升 12.36%；而相比仅优化定价策略，联合优化能带来 9.78%的收益提升。

如今大数据技术逐渐成为企业的竞争优势，然而传统的研究缺乏对企业所储存大数据资源的利用和挖掘。因此，相关理论研究也有了新的方向和发展。有学者将强化学习技术应用于收益管理领域，发现将强化学习算法如 Q-learning 算法应用在易逝产品的库存控制策略上可以更好的控制库存成本[4]。但不少使用 Q-learning 作为方法的研究难以解决大数据问题带来的维度灾难难题。在真实情况下，环境不仅十分复杂，而且常常会随着时间推移而发生改变。而深度强化学习方法无须对需求作出假设即可解决各种问题，具有更强适应性和通用性，更适应于复杂多变的真实环境。

目前深度强化学习方法较少应用于收益管理领域，本文通过深度强化学习方法研究生鲜产品联合库存控制与动态定价问题。其次，不少文献基于单一批次产品研究易逝品定价问题，同时研究库存控制与动态定价问题的文献较少。本文考虑多年龄产品共存的情况，并考虑了产品有效期随机性；除此之外，本文的状态转移函数引入了变质率作为随机变量，而时变的状态转移函数更符合实际，同时研究难度也更大。最后，本文考虑了库存状态对顾客保留价格的影响，需求是非齐次的泊松过程，以此模拟复杂多变的现实环境。在现有研究的基础上，本文的研究更贴合现实，也为企业决策提供了一定的管理依据。

收稿日期：2022-01-30；**修回日期：**2022-04-03　**基金项目：**国家自然科学基金重大研究计划：基于社会学习行为的动态定价策略研究(71871231)

作者简介：毕文杰(1967-)，男，湖南常德人，教授，博导，博士，主要研究方向为收益管理，强化学习；周玉冰(1997-)，女，湖南长沙人，硕士研究生，主要研究方向为收益管理，强化学习(zhouyub@yeah.net).

1 相关理论与方法

1.1 生鲜产品库存控制与定价策略研究

因为生鲜商品基本都具有易腐烂、易损毁的特殊自身性质，所以在易逝品库存管理与动态定价中属于一大商品类型[5]，具有一定的代表意义。有相当多论文意识到易逝品研究的重要性，专门研究易逝品的库存决策。基于变质率[6~8]、价格折扣[1, 9]、供应商延期支付[10]、有限仓储能力[11]等因素的易逝品库存控制问题受到广泛研究。早先研究者们只是将变质率设为固定的常数。而在实际生活中，易逝品变质率并非一成不变。因此研究者们逐渐考虑随时间变化的变质率，部分将其假设为服从 Weibull 分布的变量。Mishra 和 Umakanta[7]考虑时变的变质率，假设其服从 Weibull 分布，提出了一个库存控制系统。

在易逝品动态定价领域，有学者采用 WTP 模型衡量定价策略对消费者购买决策的影响[12, 13]，其中产品稀缺性也有众多学者关注[14]，Tunuguntla V 等人[15] 认为可用库存数量也会影响到顾客的保留价格，进而对消费者的支付意愿有影响。Herbon 等人[16]构建了易逝品动态定价模型，通过研究产品价格和有效期对需求的影响，分析消费者关于产品新鲜度的敏感度对定价政策的影响。

进一步地，有学者认为联合研究最优价格和最优库存将是易逝品问题研究领域新的研究方向[17]。近年来，很多文献研究了易逝品的联合库存与定价策略[18~21]。文献[18]研究了具有心理库存效应的易逝品联合定价、广告和库存控制问题，并将库存量、广告商誉、销售价格和新鲜度指数作为需求影响因素。唐跃武等人[19]基于消费者的策略性行为，研究了单阶段和两阶段定价及库存决策模型。

1.2 基于强化学习的库存控制与动态定价研究

传统库存和定价策略需要对市场需求进行假设以简化问题，与复杂多变的实际环境有一定差别。许多学者应用人工智能技术解决收益管理问题，提出了基于强化学习方法的库存控制研究，Dogan 等[22]使用强化学习算法分析了考虑多零售商竞争的零售商无限期联合订货和定价问题。Rana 和 Oliveira[23]考虑了有限期内多个相互依赖的易逝品收益管理问题，当需求是随机的且其函数形式未知时，他们使用带资格迹的 Q learning 算法来制定最佳定价政策。

强化学习方法在问题具有非常大的状态和行动空间以及未知的状态转换概率时，往往表现很差，主要是有两个问题难以解决：大的状态和行动空间导致难以存储每个状态的价值函数或最佳动作；有限的训练数据无法为每个状态提供充足的经验[24]。Q-learning 算法遇上复杂的现实问题时 Q 表将变得特别大，这将导致维度灾难问题，造成存储与检索的困难。因此强化学习由于内存复杂性、计算复杂性，以及样本复杂性局限于较小的动作空间和样本空间的低维问题[25, 26]。而深度学习正适合处理高维连续问题，学者们将强化学习与深度神经网络结合起来发展形成深度强化学习。

近年来学者们开始使用深度强化学习研究库存问题。Oroojlooyjadid 等[27]提出一种多智能体的深度强化学习算法和迁移学习方法用于啤酒游戏。他们使用一个真实世界的数据集进行测试，当其他代理使用现实行为模型时，算法表现明显优于基础库存政策。强化学习理论应用于动态定价领域也有相当多的研究。学者们使用 Q-learning[28]、策略梯度[29]等强化学习算法研究多智能体动态定价策略。Wang[30]的研究进一步使用深度强化学习来研究联合库存与定价问题，侧重于更具难度的易逝品研究。他们使用神经网络来避免维度灾难，结果表明深度强化学习模型优于没有使用神经网络的传统强化学习模型。

综上所述，目前，不少文献应用强化学习方法分别对易逝性产品的库存问题或定价问题进行了研究。但同时考虑生鲜产品的订货及库存文献较少，考虑产品变质特性的以及需求复杂性的研究更少，而这是当前企业面临的现实问题。本文的研究思路正是居于此，基于深度强化学习来建模求解生鲜产品的联合库存与定价问题，得到最大化期望总收益。

2 应用场景分析与模型构建

2.1 数学模型

目前生鲜电商往往采用“到家”或“到店+到家”经营模式，即生鲜电商平台通过在社区周边设置门店、前置仓，或者与附近商场超市、小区零售店进行合作，提供线上线下一体化服务。消费者在平台下单后，物流将快速送货到家，或者消费者选择到店自提。大型生鲜电商平台往往有着快速变化的复杂环境，这导致消费者需求具有高不确定性。同时，生鲜产品的易逝特性导致产品在存储、销售过程中发生变质、损坏、腐烂等情况，从而影响产品库存状态。因此，利用平台所储存的大量数据，不断调整和优化库存和定价策略是唯一解决方案。

本文将基于以上现实场景，考虑京东到家、盒马生鲜等大型生鲜电商的联合库存与定价问题。零售店的生鲜产品需要每日早晨由配送中心配货，零售店会根据前一日的销售情况进行预测并决定当日订货量和价格。

当顾客访问生鲜电商平台或者到达线下零售店时，会根据保留价格与实际价格决定是否购买某样产品。同时，研究发现产品的库存数量也会影响顾客心目中保留价格的设定[15]。目前，已经有部分电商平台在商品详情页显示当前库存状态，比如淘宝、亚马逊等。当顾客发现当前库存较少、产品很快就会缺货时会产生一种紧迫感，认为任何延迟可能就会错过产品，从而提高购买欲望，保留价格因而提高，因此本文假设保留价格的均值和标准差都随着库存的减少而增加。

本文引入如下假设：

a)假设配送中心可以无限量供货，产品从配货开始即进入其生命周期 l，且需要经过提前期 L 到达零售店仓库，有 $0 \leqslant L < l$；

b)本文基于 Mishra 和 Umakanta[7]的研究，以 Weibull 函数表述生鲜产品的变质率 $\theta_i(\mathrm{t})$ 作为产品的易逝特征，遵循双参数 Weibull 分布：$\theta(\mathrm{t}) = \alpha\beta \mathrm{t}^{\beta-1}$。其中 $0 \leqslant \alpha \ll 1$ 是规模参数，$\beta > 0$ 是形状参数，$0 \leqslant \theta(\mathrm{t}) \ll 1$；

c)由于生鲜产品的易逝特性，产品过期以后将停止售卖并销毁处理，残值为 0。当日产品提前售卖完也不进行补货；

d)假设售卖生鲜产品服从先进先出策略，即剩余有效期较少的产品将优先卖出；

e)假设顾客的到达遵循强度为 $\lambda(t)$ 的非齐次泊松过程[31]。异质性顾客购买概率取决于保留价格 $V(t)$，当产品价格比顾客心中的保留价格更低时，顾客将立即购买。本文还考虑顾客价格敏感系数为 $\alpha(t)$。由此，需求过程实际上可以表述为遵循 $\lambda(p,t) = \lambda(t)(1 - F(\alpha(t)p,t))$ 的非齐次泊松过程。

f)顾客保留价格 $V(t)$ 均值 $\mu(I)$ 和标准差 $\sigma(I)$ 受瞬时库存数量的影响：$\mu(I) = \mu_0 + \delta_\mu e^{-kI}$，$\sigma(I) = \sigma_0 + \delta_\sigma e^{-kI}$，其中参数 k 用于调整库存范围。

综上所述，本文联合库存与定价问题的数学模型如下：

$$\max \sum_{t=0}^{T} r^t = \sum_{t=0}^{T} \left(p^t \times n^t - c_o \times q^t - c_h \times (OI^t)^+ - c_p \times (OI^t)^- - c_f \right) \quad (1)$$
$$\text{s.t.} \quad n^t < d^t$$

其中数学符号含义如表 1 所示。

2.2 马尔可夫决策过程

上述生鲜产品的联合库存与定价是一个序列决策问题，而在多种场景下，强化学习已经被证明能够有效解决复杂序

列决策问题。其特点在于强化学习是智能系统从环境到行为映射的学习。智能体通过行为策略与环境进行交互并得到反馈。如果采取某个策略能获得正的奖励反馈，智能体将加强选择该策略的趋势，如此智能体逐步迭代学习得到最优策略。

表 1　数学符号表示

Tab. 1　Mathematical symbol meaning

数学符号	含义
r^t	t 期收益
p^t	t 期产品价格
n^t	t 期作出定价 p^t 后产品销量
c_o	单位订货成本
c_h	单位库存成本
c_p	单位缺货成本
c_f	固定成本
OI^t	t 期零售商的可用库存
s_i^t	t 期产品寿命为 i 的产品库存数量

马尔可夫决策过程是对强化学习问题的数学描述，本文将通过马尔可夫决策过程构建强化学习四元组 $\mathrm{M}=(\mathrm{S},\mathrm{A},\mathrm{P},\mathrm{R})$，应用强化学习算法求解生鲜产品的联合库存与定价问题。本文定义马尔可夫决策过程如下：

状态空间：t 期的状态变量为$(l-1)$维向量$\boldsymbol{s}^t=[s_0^t,s_1^t,\cdots,s_i^t,\cdots,s_{l-1}^t]$，$s_i^t$ 表示 t 期产品剩余有效期为 $l-i$ 的产品库存数量，L 为订货提前期，l 为产品有效期，$0\leqslant L<l$。有 $OO^t=\sum_{i=0}^{L-1}s_i^t$ 表示在途产品，当订货提前期 L 为 0 时 $OO^t=0$；$OI^t=\sum_{i=L}^{l-1}s_i^t$ 表示零售商的可用库存。如果 $OI^t>0$ 则表示当前可用库存 >0。

行动空间：t 期决策变量 $a^t=(q^t,p^t)$，q^t 为订货数量，p^t 为产品价格。订货数量将遵循 $d+x$ 规则，零售商在 $t-1$ 期观察到需求 d^{t-1}，在 t 期决定订货量为 $q^t=d^{t-1}+x^t$，即零售商将决定在上期需求基础上加减的数量。

状态转移：变质率 $\theta(\mathrm{t})$ 表示 t 时刻生鲜产品的变质特征。假设当零售商没有足够库存满足某一笔订单时，订单将会消失：当 $i<L$ 有 $s_i^t=s_{i-1}^{t-1}$，当 $L\leqslant i<l$ 时有

$$s_i^t=\theta(\mathrm{t})\left(s_{i-1}^{t-1}-\left(d^t-\sum_{j=i}^{l-1}s_j^t\right)^+\right)^+$$

奖励函数：智能体观察得到当前状态 s^t 并作出决策 a^t 以后，将得到相应的奖励 r^t，由此可以衡量出动作的价值。由状态变量 s^t 可以得知可用库存 IL^t，因此可以计算得出相应的短缺或持有成本，以及过期产品的处理成本。由决策变量中的订货数量 q^t 计算相应订货成本，以及由价格和需求得出的当期收益。由此得出奖励函数可表述为：$r^t=n^t\times p^t-c_o\times q^t-c_h\times(OI^t)^+-c_p\times(OI^t)^--\mathrm{c_d}\times(\mathrm{s}_{l-1}^t-\mathrm{n}^t)^+-c_f$. 当零售商的可用库存大于需求时，销量 $n^t=d^t$；当零售商缺货时，未满足的订单不再延续到下一期，$n^t=OI^t$。

t 期事件发生顺序如下：

a)期初更新 t 期的状态为 S^t；

b)智能体作出决策 a^t，决定订货数量 q^t 和产品价格 p^t，产品将在 L 期后到达。当 $L=0$，产品将马上到达；

c)需求 d^t 到达，在零售商缺货的情况下，未满足的订单将消失。而交付订单后，余下有效期内库存将转移至下个周期，过期产品将被处理；

d) t 期末结算该期收益与成本，智能体收到奖励 R^t，并且更新状态至 S^{t+1}。

3　生鲜产品联合定价和库存控制算法设计

强化学习的发展历史最开始可以追溯到 Bellman 提出的贝尔曼条件以及马尔可夫决策过程，并以贝尔曼期望方程描述了状态价值函数与动作价值函数之间的关系，这正是强化学习方法的理论基础。但直接求解贝尔曼期望方程是不切实际的，强化学习逐渐发展出基于价值的强化学习方法(Value-Based RL)、基于概率的强化学习方法(Policy-Based RL)。前者主要代表算法有基于在线更新的 SARSA 算法以及离线学习策略的 Q-learning 算法；后者有策略迭代算法(Policy Gradients)，它针对连续动作空间，直接输出下一动作概率。除此之外，以上两类算法合并形成了行为-评判算法(actor-critic)，该算法结合两类算法的优点，使用 Critic 学习奖惩机制，Actor 输出动作概率。

强化学习的目的是通过学习策略获得最优收益。作为值函数迭代强化学习方法，Q-learning 算法将通过 Q 值表记录状态-行动对的值，并基于 bellman 方程更新 Q 表：

$$\mathrm{Q(S_t,A_t)}\leftarrow\mathrm{Q(S_t,A_t)}+\alpha\left[\mathrm{R_{t+1}}+\gamma\max_{\mathrm{a}}\mathrm{Q(S_{t+1},a)}-\mathrm{Q(S_t,A_t)}\right]\tag{2}$$

Q-learning 有明显的局限性，当状态空间或者动作空间特别大时，Q-learning 算法难以建立和维护一个巨大的 Q 值表，因此难以记录所有的状态和动作。为了解决维度灾难问题，Mnih 等[32]将 Q-learning 算法与深度学习结合，扩展成 Deep Q-network(DQN)，它使用神经网络来近似 Q 值表，即以函数来表示状态-动作对 Q(s,a)：$\mathrm{Q(s,a)}\leftarrow\mathrm{f(s,a,w)}$

DQN 可以由当前奖励和下一状态的 Q 估计值来估计 Q 目标值，并最小化 Q 估计值和 Q 目标值之间的差异，以此更新神经网络的参数，其损失函数采用估计值和目标值的均方差：

$$\mathrm{L(w)}=\mathbb{E}\left[\left(\mathrm{r}+\gamma\max_{\mathrm{a'}}\mathrm{Q(s',a',w)}-\mathrm{Q(s,a,w)}\right)^2\right]\tag{3}$$

同时，强化学习采集的训练数据之间往往存在相关性，这会导致神经网络不稳定，DQN 使用了经验回放机制(replay buffer)来避免该问题。即采用经验池 $\mathcal{D}$ 存储每步探索的数据 $\{\mathrm{S,A,R,S'},isdone\}$，从中取样并根据取样数据计算 Q 目标值，更新当前神经网络参数。

为了解决 DQN 的过估计等问题，优化算法的性能表现，学者们逐渐提出了双重 DQN 算法(Double DQN)、优先级经验回放(DQN with Prioritized Replay)以及对偶 DQN(Dueling DQN)等方法。在对偶 DQN 算法中，学者修改了 DQN 的神经网络架构，大幅提高了算法的效率。相比于 DQN 使用神经网络直接输出各个动作的 Q 值，对偶 DQN 定义了一个优势函数 $A(s,a)=Q(s,a)-V(s,a)$

Q 值将由状态价值估计值 $V(s,a;w)$ 和优势函数估计值 $A(s,a;w)$ 确定：

$$Q(s,a;w)=V(s;w)+A(s,a;w)\tag{4}$$

综上所述，在本文的联合定价和库存控制问题中，智能体的状态设置为 $\boldsymbol{s}^t=[s_0^t,s_1^t,\cdots,s_i^t,\cdots,s_{l-1}^t]$，表示当期的库存状态。本文采用经验回放与固定目标网络机制，设置两个结构相同但参数不同的神经网络，即评估网络和目标网络。同时为 Q 值定义优势函数，用以优化算法的性能。神经网络的具体细节设置将在下一章节进行详细描述。每期智能体将按照贪婪策略作出决策，从订货和定价的动作空间中选择某一动作。智能体采取行动后，将收到环境的反馈，从而估计出目标 Q 值，并更新评估神经网络参数。经过固定数量的步骤后，将评估网络的参数值赋给目标网络。本文生鲜产品库存控制和定价具体算法如下所示。

算法　生鲜库存控制和定价算法

输入：环境。

输出：动作价值函数 $q(s,a)$。

a)(初始化)初始化评估网络 q 的参数 θ 和目标网络 $\hat{q}$ 的参数 $\hat{\theta}$；初始化经验池 $\mathcal{D}$，其大小设置为 N。

逐回合执行：

b)初始化环境和状态 S_0；

c)从 $t=1$ 到 $t=T$，执行以下操作：

(a)(贪婪策略)利用概率 ε，从 $\mathcal{A}(s)$ 中等概率随机选择一个动作 A；否则，$A = arg\max_a q(S_t, a)$；

(b)执行动作 A，观测环境得到奖励 R 和下一状态 S'

(c)(经验存储)将四元组 (S, A, R, S') 存入经验池 $\mathcal{D}$ 中；

(d)(经验回放)从经验池 $\mathcal{D}$ 中采样出一批数据 $(S_i, A_i, R_i, S_{i'})$，$\forall i = 1,\ldots,N$；

(e)计算回报的估计值：

$$y_i = \begin{cases} R_i & i+1\text{期回合终止} \\ R_i + \gamma \max_{A'} \hat{q}(S_i', A'; \hat{\theta}) & \text{其他} \end{cases};$$

(f)更新动作价值函数逼近的神经网络参数 θ；

(g)更新状态变量 $S \leftarrow S'$

(h)每隔 C 步更新目标网络的参数 $\hat{\theta} \leftarrow \theta$。

4　仿真实验

实验将设定具体的实验参数，通过深度强化学习算法与设计好的模拟环境进行交互，得到具体的数值结果。以此分析算法在仿真应用环境中的表现结果，并判断算法能否应用于真实环境。实验采用控制变量法，比较学习率（learning rate）、gamma 值参数对实验结果的影响。

根据上述模型与算法分析，本文首先对算法的神经网络进行设置。算法设有两个结构相同的神经网络，其参数分别为 θ 和 θ'。每个神经网络有两个隐藏层，并使用 ReLU 激活函数。设置经验池的容量大小 N 为 10000，每回合将随机从中采样。更新目标网络的间隔步数 C 设置为 300 步。

实验的通用参数设置为

a)生鲜产品生命周期 l 设置为 $\{4,5,6\}$，提前期 L 为 $\{0,1,2\}$；

b)产品价格集合采用折扣率形式的离散定价集合，$\text{discount} = \{0.1, 0.2, \cdots, 0.9, 1.0\}$，原始定价 $p_{\text{base}} = 20$；

c)产品订购数量采取 d+x 形式，设置 x 取值范围为 $\{-20, \cdots, 20\}$；

d)对于 ϵ-greedy 策略，设置探索值 ϵ 在学习过程中将逐步衰减到阈值，初始时 $\epsilon_{\text{init}} = 0.9$，在迭代中线性递减直到 $\epsilon_{\text{end}} = 0.1$：$\epsilon = \epsilon - \dfrac{\epsilon_{\text{init}} - \epsilon_{\text{end}}}{\text{episodes} \times \beta}$。

4.1　深度学习超参数的选择与测试

由于算法引入了深度神经网络，并且模型的超参数一般需要手动设置，不同于由数据估计而来的参数，超参数的选择将直接影响到智能体训练的稳定性和收敛性，关系到训练效果的好坏，因此如何选取超参数，保证策略的有效性是强化学习在实际应用中的关键问题。本文将针对学习率和 gamma 值超参数作出比较，分析其对训练效果的影响。

a)在其他参数相同的情况下，分别测试了学习率 α=0.005、α=0.001、α=0.0005 的情况。实验结果如图 1 所示。学习率是深度学习模型训练中非常重要的参数，关系到神经网络参数的更新程度，进而影响模型的收敛。学习率参数值设置得过高或过低都将导致模型性能表现不佳：学习率过高将可能导致模型不收敛，当学习率选取 0.005 时算法前期震荡幅度非常大，后期表现也不如选取学习率为 0.001 的算法；学习率过低则将导致算法训练时间过长，需要更多时间才能收敛。

b)固定其他参数，分别测试了 gamma 值 γ=0.9、γ=0.95、γ=0.99 的情况，实验结果如图 2 所示。对于 gamma 值而言，参数设置越高，智能体将越注重未来总体收益，难于关注眼前的短期收益，因此训练将困难、缓慢。

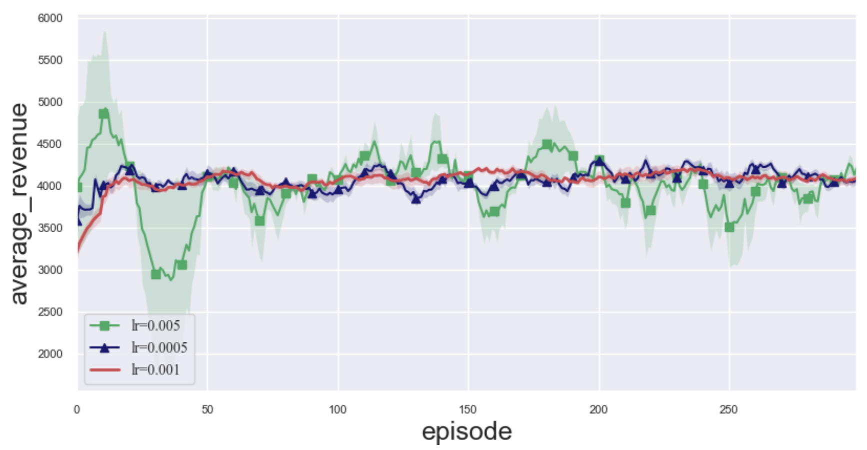


图 1　学习率 α=0.005、α=0.001、α=0.0005

Fig. 1　Learning rate at α=0.005, α=0.001, α=0.0005



图 2　gamma 值 γ=0.9、γ=0.95、γ=0.99

Fig. 2　gamma value at γ=0.9, γ=0.95, γ=0.99

4.2　深度强化学习算法性能实验

强化学习方法应用于收益管理领域已经有一定研究，Q-learning 算法与 SARSA 算法应用较为广泛[33, 34]；深度强化学习方法中的 DDPG 算法在电子商务领域也得到应用[35]。本文将基于上述参数设置，与以下基准模型进行对比实验：

a)DDPG 算法是将深度神经网络融入确定性行为策略的策略学习方法，本文设置 DDPG 的两个 Actor 网络和两个 Critic 网络都为两层全连接层，激活函数为 Relu 函数，学习率 α=0.001，gamma 值 γ=0.95，探索值 ε 设置为前文衰减策略；

b)表格式强化学习方法：SRASA 算法通过同策时序差分更新求解最优策略，Q learning 则是异策算法，采用不同的方法更新最优动作价值估计。本文设置 SRASA 算法与 Q learning 算法的学习率、gamma 值等超参数均与上述一致，算法结果如图 3 所示。

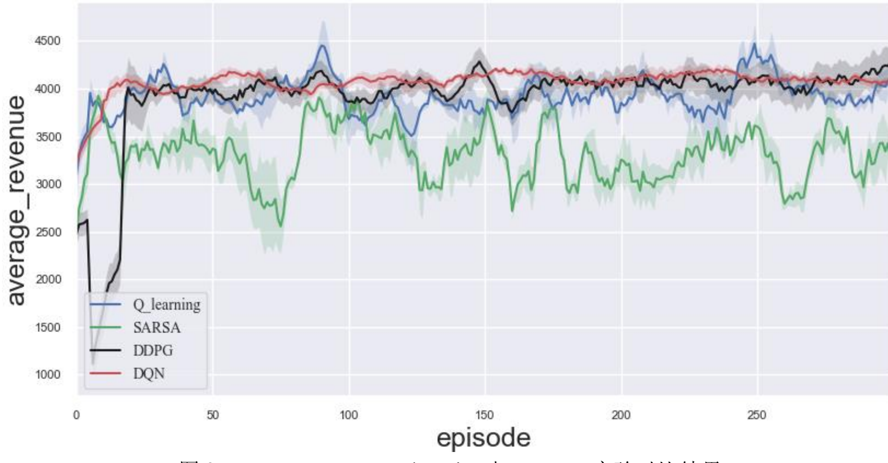


图 3　DQN、DDPG、Q-learning 与 SARSA 实验对比结果

Fig. 3　Results for DQN、DDPG、Q-learning and SARSA

由表 2 可见，DQN 算法的收益表现最佳，其次是 DDPG 算法和 Q-learning 算法，SARSA 算法的收益最低。而在稳定性方面 DQN 算法也优于其他算法，相较于 DQN 算法，DDPG 和 Q-learning 算法都比较震荡。

表 2　模型收益表现对比

Tab. 2　Performance comparison between models

模型	中位数	平均收益	收益上界	收益下界
DQN	4239	4228	5854	1021
DDPG	4276	4186	6248	1209
Q-learning	4032	4032	5248	2920
SARSA	3476	3480	6627	425

现实中的需求变化极其复杂，各种因素导致需求呈现随机性强、非稳态的波动变化。DQN 库存控制与动态定价算法能够解决维度灾难问题，也能为企业提供近似最优的订货和定价策略。由此可见，基于 DQN 方法的联合库存控制与动态定价模型具有非常广泛的应用价值。

5　结束语

目前生鲜产品需求逐年递增，市场规模也相应扩大，对于零售商而言如何合理控制库存与定价是非常重要的决策问题。本文研究了生鲜产品的联合库存控制与动态定价问题，在需求不断变化的情况下通过深度强化学习算法探索生鲜产品的最优订货量和最优定价，从而达到企业收益最大化的目的。

首先，本文引入了时变的变质率作为状态转移的一部分。随着时间推移，生鲜产品将或多或少损坏或腐烂，通过变化的变质率描述该现象更符合实际。而强化学习方法对于未知的状态转移函数表现不佳，目前基于强化学习的收益管理研究很少涉及到变化的状态转移方程，本文通过深度强化学习方法弥补了该方面的不足。

除此之外，当顾客感知到产品库存较少，商品即将缺货时的紧迫感会提高其购买欲望，从而影响到产品需求。考虑到这一点，零售商可以战略性地决定库存和定价策略，以使其收益最大化。基于受库存影响的顾客支付意愿，本文研究了零售商在短销售期限内销售生鲜产品的库存和定价策略。目前学术界少有相关研究，本文研究成果对易逝品收益管理领域有所贡献。

在生鲜产品联合库存控制与动态定价问题中，本文通过设计的深度强化学习算法，根据当前可用库存来动态调整价格和订货量，使预期利润最大化。本文只关注单一代理人的联合定价和库存控制问题。未来可以在考虑竞争情况下的库存控制和动态定价方面进行深入的研究。

参考文献：

[1] 但斌，陈军，吴庆. 基于多级折扣价格的易逝品订货策略研究 [J]. 中国管理科学, 2006(3): 38-44. (Dan Bin, Chen Jun, Wu Qing. Optimal ordering policy for perishable product with progressive price discounts [J]. Chinese Journal of Management Science, 2006(3): 38-44.)

[2] 赵泉午，熊中楷，林娅，等. 基于电子市场的易逝品两级供应链供需博弈分析 [J]. 中国管理科学, 2004(3): 92-7. (Zhao Quanwu, Xiong Zhongkai, Lin Ya, *et al.* Game analysis of two-stage supply chain for perishable goods under e-marketplace [J]. Chinese Journal of Management Science, 2004(3): 92-7.)

[3] Fang F, Nguyen T-D, Currie C S. Joint pricing and inventory decisions for substitutable and perishable products under demand uncertainty [J]. European Journal of Operational Research, 2021, 293 (2): 594-602.

[4] Kara A, Dogan I. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems [J]. Expert Systems with Applications, 2018, 91: 150-8.

[5] Li R, Lan H, Mawhinney J R. A review on deteriorating inventory study [J]. Journal of Service Science and Management, 2010, 3 (01): 117.

[6] Ferguson M, Ketzenberg M E. Information sharing to improve retail product freshness of perishables [J]. Production and Operations Management, 2006, 15 (1): 57-73.

[7] Mishra U. An EOQ model with time dependent Weibull deterioration, quadratic demand and partial backlogging [J]. International Journal of Applied and Computational Mathematics, 2016, 2 (4): 545-63.

[8] 李业梅，黄少安. 基于 EOQ 模型的非瞬时变质食品提前支付订货策略研究 [J]. 中国管理科学: 1-13. (Li Yemei, Huang Shaoan. Advance payment strategy of non-instantaneous food based on EOQ model [J]. Chinese Journal of Management Science, 1-13.)

[9] Viswanathan S, Wang Q. Discount pricing decisions in distribution channels with price-sensitive demand [J]. European Journal of

Operational Research, 2003, 149 (3): 571-87.

[10] 徐贤浩, 王倩, 曾款, 等. 延迟支付条件下易逝品的最优订货决策研究 [J]. 中国管理科学, 2021, 29 (02): 108-16. (Xu Xianhao, Wang Qian, Zeng Kuan, *et al.* Study on the optimal ordering policy of perishable products with delayed payment [J]. Chinese Journal of Management Science, 2021, 29 (02): 108-16.)

[11] 靖富营, 潘杨. 仓储能力约束和缺货下两易逝品联合采购动态批量决策 [J]. 系统工程, 2018, 36 (7): 47-54. (Jing Fuying, Pan Yang. A two product dynamic lot size model with perishable inventory and joint ordering under bounded inventory [J]. Systems Engineering, 2018, 36 (7): 47-54.)

[12] Wertenbroch K, Skiera B. Measuring consumers'willingness to pay at the point of purchase [J]. Journal of marketing research, 2002, 39 (2): 228-41.

[13] Kalish S. A new product adoption model with price, advertising, and uncertainty [J]. Management science, 1985, 31 (12): 1569-85.

[14] Zhu M, Ratner R K. Scarcity polarizes preferences: The impact on choice among multiple items in a product class [J]. Journal of Marketing Research, 2015, 52 (1): 13-26.

[15] Tunuguntla V, Basu P, Rakshit K, *et al.* Sponsored search advertising and dynamic pricing for perishable products under inventory-linked customer willingness to pay [J]. European Journal of Operational Research, 2019, 276 (1): 119-32.

[16] Herbon A, Khmelnitsky E. Optimal dynamic pricing and ordering of a perishable product under additive effects of price and time on demand [J]. European Journal of Operational Research, 2017, 260 (2): 546-56.

[17] Dye C-Y, Hsieh T-P, Ouyang L-Y. Determining optimal selling price and lot size with a varying rate of deterioration and exponential partial backlogging [J]. European Journal of Operational Research, 2007, 181 (2): 668-78.

[18] Dye C-Y. Optimal joint dynamic pricing, advertising and inventory control model for perishable items with psychic stock effect [J]. European Journal of Operational Research, 2020, 283 (2): 576-87.

[19] 唐跃武, 范体军, 刘莎. 考虑策略性消费者的生鲜农产品定价和库存决策 [J]. 中国管理科学, 2018, 26 (11): 105-13. (Tang Yuewu, Fan Tijun, Liu Sha. Pricing and inventory decision-making for fresh agricultural products with strategic consumers [J]. Chinese Journal of Management Science, 2018, 26 (11): 105-13.)

[20] 王淑云, 姜樱梅, 牟进进. 基于新鲜度的冷链一体化库存与定价联合决策 [J]. 中国管理科学, 2018, 26 (7): 132-41. (Wang Shuyun, Jiang Yingmei, Mou Jinjin. Inventory and pricing decision of an integrated cold chain based on freshness [J]. Chinese Journal of Management Science, 2018, 26 (7): 132-41.)

[21] 曹裕, 易超群, 万光羽. 易逝品随机生产库存模型动态定价, 服务水平和生产控制策略 [J]. 系统工程理论与实践, 2018, 38 (7): 1717-31. (Cao Yu, Yi Chaoqun, Wan Guangyu. Dynamic pricing, service and production control strategy of stochastic production-inventory models with perishable products [J]. Systems Engineering—Theory & Practice, 2018, 38 (7): 1717-31.)

[22] Dogan I, Guener A R. A reinforcement learning approach to competitive ordering and pricing problem [J]. Expert Systems, 2015, 32 (1): 39-48.

[23] Rana R, Oliveira F S. Dynamic pricing policies for interdependent perishable products or services using reinforcement learning [J]. Expert Systems with Applications, 2015, 42 (1): 426-36.

[24] Zarandi M H F, Moosavi S V, Zarinbal M. A fuzzy reinforcement learning algorithm for inventory control in supply chains [J]. The International Journal of Advanced Manufacturing Technology, 2013, 65 (1-4): 557-69.

[25] Arulkumaran K, Deisenroth M P, Brundage M, *et al.* A brief survey of deep reinforcement learning [J]. arXiv preprint arXiv: 170805866, 2017.

[26] Strehl A L, Li L, Wiewiora E, *et al.* PAC model-free reinforcement learning [C]// Proceedings of the 23rd international conference on Machine learning. 2006: 881-888.

[27] Oroojlooyjadid A, Nazari M, Snyder L V, *et al.* A Deep Q-Network for the Beer Game: Deep Reinforcement Learning for Inventory Optimization [J]. Manufacturing & Service Operations Management, 2021.

[28] Jintian W, Lei Z. Application of reinforcement learning in dynamic pricing algorithms [C]// 2009 IEEE International Conference on Automation and Logistics. IEEE, 2009: 419-423.

[29] Könönen V. Dynamic pricing based on asymmetric multiagent reinforcement learning [J]. International journal of intelligent systems, 2006, 21 (1): 73-98.

[30] Wang R, Gan X, Li Q, *et al.* Solving a Joint Pricing and Inventory Control Problem for Perishables via Deep Reinforcement Learning [J]. Complexity, 2021, 2021.

[31] Zhao W, Zheng Y-S. Optimal dynamic pricing for perishable assets with nonhomogeneous demand [J]. Management science, 2000, 46 (3): 375-88.

[32] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529-533.

[33] 王欣, 王芳. 基于强化学习的动态定价策略研究综述 [J]. 计算机应用与软件, 2019, 36 (12): 1-6. (Wang Xin, Wang Fang. A review of dynamic pricing strategy based on reinforcement learning [J]. Computer Applications and Software, 2019, 36 (12): 1-6.)

[34] Gašperov B, Begušić S, Posedel Šimović P, *et al.* Reinforcement Learning Approaches to Optimal Market Making [J]. Mathematics, 2021, 9 (21): 2689.

[35] Mosavi A, Faghan Y, Ghamisi P, *et al.* Comprehensive Review of Deep Reinforcement Learning Methods and Applications in Economics [J]. Mathematics, 2020, 8 (10): 1640.